AP Statistics Chapter 7 Sampling Distributions

Statistical inference looks at how often would this method give a correct answer if it was used many many times. Statistical inference works best when we produce data by random sampling or randomized comparative experiments. The reason is that when we use chance to choose respondents or assign subjects, the laws of probability answer the question "what would happen if we did this many times? We begin by looking at sampling distributions and simple ways to describe said distributions.

Basic Terms

! A parameter is a number that describes the population. In statistical practice, the value of a parameter is seldom known.

! A statistic is a number than can be computed from the sample data without making use of any unknown parameters. <u>Typically, we use a statistic to estimate an unknown parameter.</u>

EXAMPLE: Are attitudes toward shopping changing? Sample surveys show that fewer people enjoy shopping than in the past. A recent survey asked a nationwide random sample of 2500 adults if they agreed or disagreed that "I like buying new clothes, but shopping is often frustrating and time-consuming". Of the respondents, 1560 or 66% said they agreed. The number 66% is a STATISTIC. The population that the poll wants to draw conclusions about is all U.S. residents age 18 and over. The parameter of interest is the percent of all adult U.S. residents who would have said Agree if asked the same questions. We don't know the value of the parameter.

Parameter or Statistic???

A carload lot of ball bearings has mean diameter 2.5003 centimeters. This is within the specifications for acceptance of this lot by the purchaser. By chance, an inspector chooses 100 bearings from the lot that have mean diameter 2.5009 cm. Because this is outside the specified limits, the lost is mistakenly rejected.

A telemarketing firm in Oak Ridge uses a device that dials residential telephone numbers in the city at random. Of the first 100 numbers dialed 48% are unlisted. This not surprising because 52% of all Oak Ridge residential phones are unlisted.

A researcher carries out a randomized comparative experiment with young rats to investigate the effects of a toxic compound in food. She feeds the control group a normal diet. The experimental group receives a diet with 2500 parts per million of the toxic material. After 8 weeks, the mean weight gain is 335 grams for the control group and 289 grams for the experimental group.

VARIABILITY

Many times we are interested in the proportion of the population that has a particular characteristic (e.g., proportion that finds clothes shopping frustrating). The population proportion (called p) is a parameter. Our poll in the previous example found that 1650 out of 2500 randomly selected adults agreed with the statement that shopping is often frustrating. The proportion of the sample who agreed was

$$\hat{p} = \frac{1650}{2500} = 0.66$$

The sample proportion p-hat is a statistic.

We use the value of p-hat to estimate the value p of the population.

But wait a minute. How can a statistic based upon a sample of 2500 be a reasonable estimate of a parameter which is characteristic of >180 million American adults. Certainly, a second random sample taken at the same time would choose different people and probably produce a different value of p-hat. This basic fact is called SAMPLING VARIABILITY: the value of a statistic varies in repeated random sampling.

But rejoice!!! Sampling variability is not fatal. We simply ask the question what would happen if we took many samples? To do this we

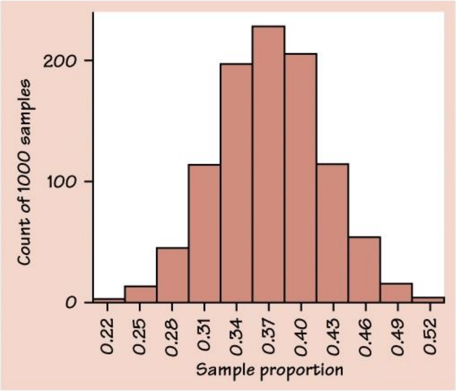
! Take a large number of samples from the same population
! Calculate the sample proportion p-hat for each sample
! Make a histogram of the values of p-hat
! Examine the distribution displayed in the histogram for overall pattern, center and spread and outliers or other deviations.

SAMPLING DISTRIBUTION The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Describing Sampling Distributions

Advertisers on TV are very interested in how many viewers watch a particular television show. According to the 2001 Nielsen ratings, Survivor II was one of the most-watched television shows in the United States during every week that it aired. Suppose that the true proportions of U.S. adults who watched Survivor II is p = 0.37.

The following figure shows the results of drawing 1000 SRSs of size n = 100 from a population with p = 0.37



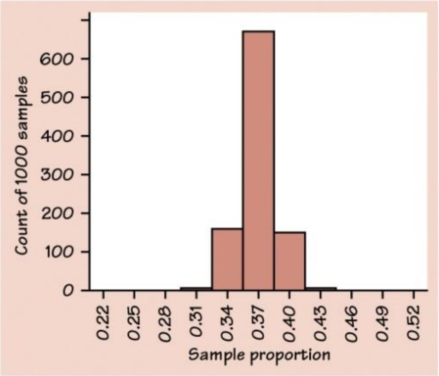
Looking at the sampling distribution we see that

- The overall shape of the distribution is symmetric and approximately normal
- The center of the distribution is very close to the true value of p = 0.37 (actual mean and median are 0.372 and 0.37, respectively)
- The values of p-hat have a large spread. They range from 0.22 to 0.54. The standard deviation is calculated to be 0.05.
- There are no outliers or other important deviations from the overall pattern

Note---When randomization is used in a design for producing data, statistics computed from the data have a definite pattern of behavior over many repetitions, even though the result of a single repetition is uncertain. HOWEVER, haphazard sampling does not give such regular and predictable results.

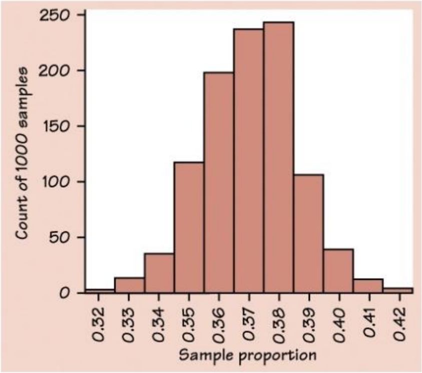
The fact that statistics from random samples have definite sampling distributions allows a more careful answer to the question of how trustworthy a statistic is as an estimate of a parameter.

If we repeat the sample survey for people who watched Survivor II, but this time with a sample size of 2500 we get the following figure



The scale is the same as in the previous figure

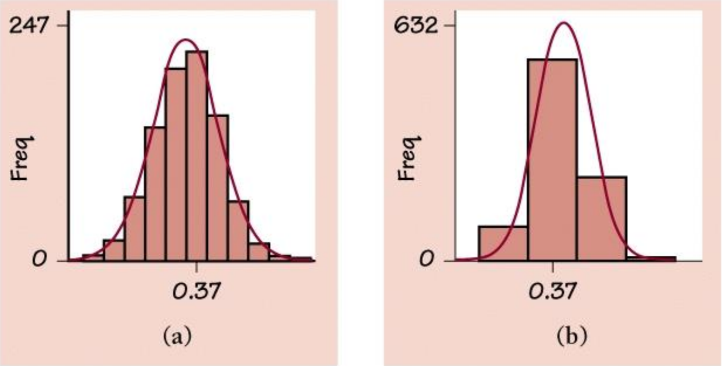
Expanding the scale to better identify its shape



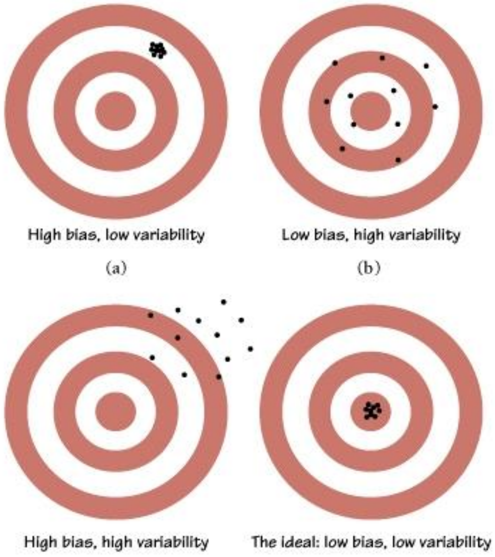
Looking at the distribution we see that

- The pattern is symmetric and approximately normal
- The center of the distribution is close to 0.37 (actual mean and median are 0.3693 and 0.37, respectively)
- The spread is much less with the range of the values being 0.321 to 0.421. The standard deviation is 0.016.

The figure below shows two sampling distributions of p-hat for sample size 100 and 2500. Both distributions are approximately normal, so we have also drawn normal curves for both.

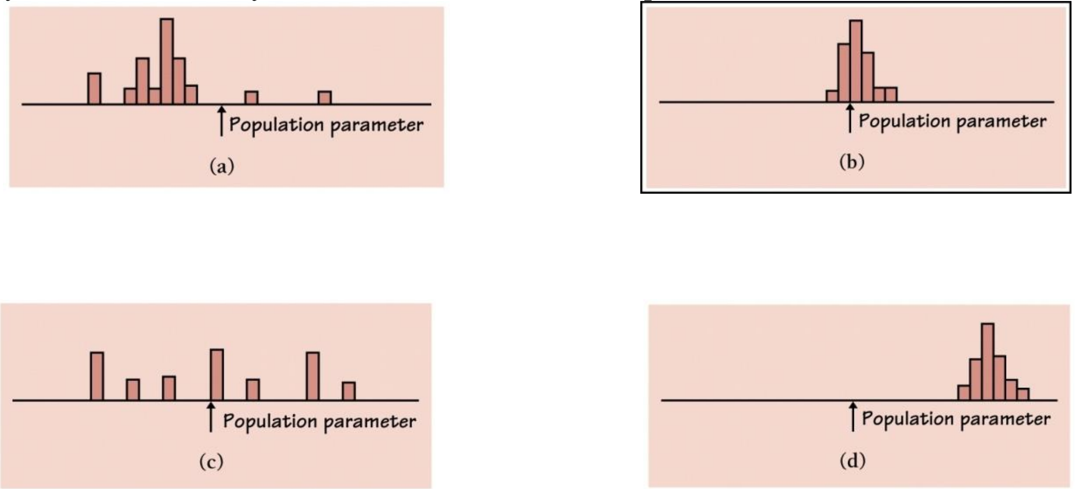


In addressing the question of how trustworthy is the sample proportion p-hat as an estimator of the population proportion p in each case we now look at both bias and variability


High bias, low variability
(a)
Low bias, high variability
(b)
High bias, high variability
(c)
The ideal: low bias, low variability
(d)

A statistic used to estimate a parameter is UNBIASED if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

Note that individual values may fall below or above the true value but there is no systematic tendency to over/under-estimate the parameter.



The variability of a statistic is described by the spread of its sampling distribution.
This spread is determined by the sampling design and the size of the sample.
LARGER SAMPLES GIVE SMALLER SPREAD.

IMPORTANT NOTE---As long as the population is much larger than the sample (say at least 10 times as large), the spread of the sampling distribution is approximately the same for any population size (for a given sample size, n)

That is a statistic from an SRS of size 2500 from the more than 250,000,000 residents of the U.S. is just as precise as a statistic from an SRS of size 2500 from the 740,000 inhabitants of San Francisco. This is great news for the designers of surveys. But remember, to obtain equally trustworthy results both must use the same sample size.

EXAMPLE The Internal Revenue Service plans to exam an SRS of individual federal income tax returns from each state. One variable of interest is the proportion of returns claiming itemized deductions. The total number of tax returns in a state varies from almost 14 million in California to fewer than 210,000 in Wyoming.

a. Will the sampling variability of the sample proportion change from state to state if an SRS of 2000 tax returns is selected in each state? Explain.

b. Will the sampling variability of the sample proportion change from state to state if an SRS of 1% of all tax returns is selected in each state? Explain.

In Part a, sample size is constant and thus so is variability.

In Part b, the larger the sample size the smaller the variability.

Sampling Distribution of sample proportions

So far we have seen that the sampling distribution of a sample proportion (p-hat) can give a pretty good estimate of the actual population proportion (parameter p).

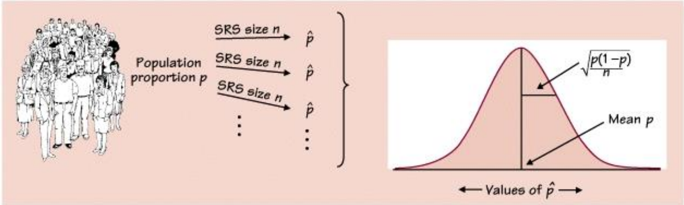
Here are the facts:

- Choose an SRS/Random sample of size n from a large population with population proportion p having some characteristic of interest.
- Let p-hat be the proportion of the sample having that characteristic.
- Then the sampling distribution of p-hat is approximately normal and is closer to a normal distribution when the sample size n is large.

The mean of the entire (or idealized) sampling distribution of p-hat is exactly p.

The standard deviation of the sampling distribution of p-hat is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Because the mean of the sampling distribution of p-hat is always equal to the parameter p, the sample proportion p-hat is an unbiased estimator of p (when sample is random)



The standard deviation of p-hat gets smaller as the sample size n increases because n appears in the denominator of the formula for the standard deviation. That is, p-hat is less variable in larger samples. Note that the sample size n is under the square root sign so to cut the standard deviation in half, we must take a sample four times as large, not just twice as large.

The formula for the standard deviation of p-hat does not apply when the sample is a large part of the population. For example you cannot use this recipe if you choose an SRS of 50 of the 100 people in a class. In practice, we take a sample only when the population is large. Otherwise, we examine the entire population.

- Check for randomness in sample selected first. That gives you an unbiased estimator.
- Use the recipe for the standard deviation of p-hat only when the population is at least 10 times as large as the sample. MAKE SURE YOU VERIFY THIS WHEN YOU USE THE RECIPE!!!

- Use the normal approximation to the sampling distribution of p-hat for values of n and p that satisfy

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

Again, VERIFY THIS WHEN YOU USE IT!!!!!

EXAMPLE You ask an SRS of 1500 first year college students whether they applied for admission to any other college. There are over 1.7 million first-year college students. It is known that 35% of all first-year students applied to colleges besides the one they are attending. What is the probability that your sample will give a result within 2 percentage points of this true value??

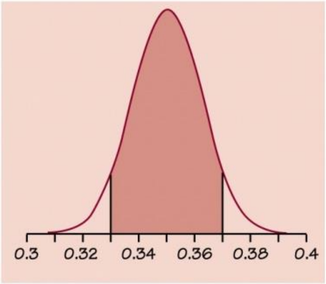
Verification checks

1. SRS stated so p-hat is an unbiased estimator for p.
2. Population of 1.7 million is greater than 10 times sample size of 1500 so:

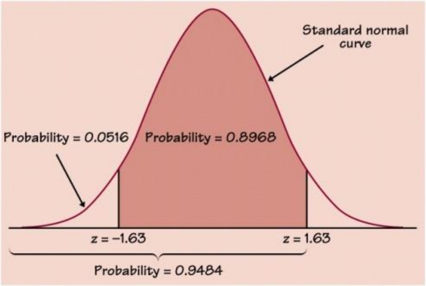
The standard deviation is calculated to be $\sigma_{\hat{p}} = \sqrt{\frac{(0.35)(1-0.35)}{1500}} = 0.0123$

3. $np \geq 10$ $(1500)(0.35) = 525 \geq 10$
$n(1-p) \geq 10$ $(1500)(1-0.35) = 975 \geq 10$

Since the sample is large enough, tha sampling distribution of the proportion of students that applied to other colleges is approximately normal.



The normal approximation to the sampling distribution of p-hat



The standard normal curve

EXAMPLE

One way of checking the effect of undercoverage, nonresponse and other sources of error in a sample survey is to compare the sample with known facts about the population. About 11% of American adults are black. The proportion p-hat of blacks in an SRS of 1500 adults should therefore be close to 0.11. It is very unlikely to be exactly 0.11 because of sample variability. If a national sample of 1500 Americans contains only 9.2% blacks, should we suspect that the sampling procedure is somehow underrepresenting blacks?

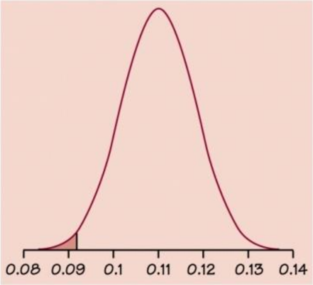
To answer this question find the probability that a sample contains no more than 9.2% blacks when the population is truly 11% black. First do a verification check for using the equations and the normal approximation

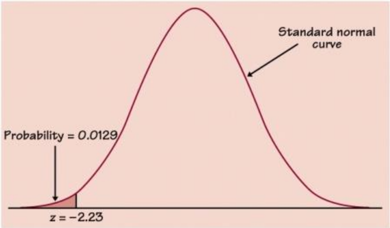
- SRS is given so: The mean of p-hat is 0.11
- Actual population of American adults is greater than 10 times 1500 so:

the standard deviation is: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.11)(1-0.11)}{1500}} = 0.0081$

- np ≥ 10 (1500)(0.11) = 165 ≥ 10 ; n(1-p) ≥ 10 (1500)(1-0.11) = 1335 ≥ 10
  Since the sample is large enough, tha sampling distribution of the proportion of blacks in the population is approximately normal.

AP Statistics Chapter 7 Sampling Distributions





The Standard Normal Distribution

We have been discussing sample proportions. Sample proportions arise most often when we are interested in a categorical variable, like are you frustrated with shopping, do you like red, or do you kiss on the first date. In these cases we would answer the question by saying that so many percent of the population likes red or the like. However, when we record quantitative variables such as blood pressure, heart rate or height, then we can calculate the actual numerical average of any data set directly. Indeed we often use average, mean or expected value to describe a data set because

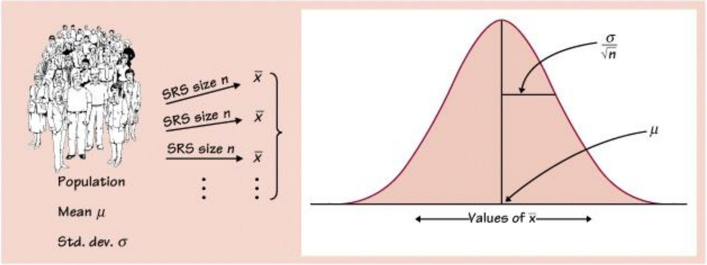
! Averages are less variable than individual observations
! Distributions of averages are more normal than distributions of individual observations

The mean and standard deviation of a population are parameters and are designed by the Greek Letters μ for the mean and σ for the standard deviation.

The mean and standard deviation calculated from sample data are statistics. We write the sample mean as x-bar and the sample standard deviation as s.

Suppose that x-bar is the mean of an SRS of size n drawn from a large population with mean μ and standard deviation σ. Then the mean of the sampling distribution of x-bar is $\boldsymbol{\mu_{\bar{x}} = \mu}$ and its standard deviation $\boldsymbol{\sigma_{\bar{x}} = \sigma/\sqrt{n}}$

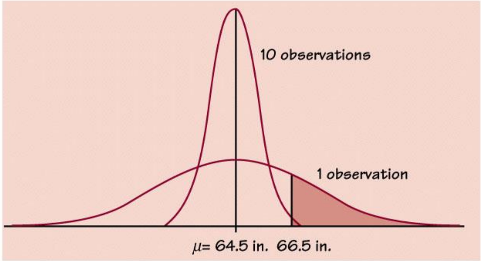
The behavior of x-bar in repeated samples is similar to that of the sample proportion
! The sample mean x-bar is an unbiased estimator of the population mean μ
! The values of x-bar are less spread out for larger samples. Their standard deviation decreases at the rate √n so you must take a sample four times as large to cut the standard deviation of x-bar in half.
! You should only use the recipe σ/(√n) for the standard deviation of x-bar when the population is at least 10 times as large as the sample. VERIFY. VERIFY. VERIFY.



EXAMPLE The height of young women varies approximately according to the N(64.5, 2.5) distribution. This is a population distribution with μ = 64.5 and σ = 2.5. If we choose one young woman at random, the heights we get in repeated choices follow this distribution. That is, the distribution of the population is also the distribution of one observation chosen at random. So we can think of the population distribution as a distribution of probabilities, just like a sampling distribution.

If we chose an SRS of 10 young women and determined the average height (x-bar); the sampling distribution of many repeated selections would have a mean of 64.5 and a standard deviation of σ/($\sqrt{n}$) = 2.5/($\sqrt{10}$) = 0.79.

Determine the probability that you select a sample of 10 young women with average height $\geq$ 66.5 inches



The sampling distribution of the mean height x-bar for samples of 10 young women compared with the distribution of the height of a single woman chosen at random

REAL WORLD STUFF---The fact that averages of several observations are less variable than individual observations and approximates the true average of all possible observations is important to the way the real world does business. For example, it is common practice to repeat a careful measurement several times and report the average of the results. Think of the results of n repeated measurements as an SRS from the population of outcomes we would get if we repeated the measurement forever. The average of the n results (the sample mean x-bar) is less variable than a single measurement and a better estimator of the true average.

EXAMPLE An automatic grinding machine in an auto parts plant prepares axles with a target diameter μ = 40.125 mm. The machine has some variability, so the standard deviation of the diameters is σ= 0.002 mm. The machine operator inspects a sample of 4 axles each hour for quality control purposes and records the sample mean diameter. What will be the mean and standard deviation of the numbers recorded? Do your results depend on whether or not the axle diameters have a normal distribution?

Assuming the number of axles per hour is greater than 40 then the mean is 40.125 and the standard deviation is $\frac{\sigma}{\sqrt{n}} = \frac{0.002}{\sqrt{4}} = 0.001$. The results only depend on the fact that the population (axles per hour) is large (10 times or greater) compared to the sample.

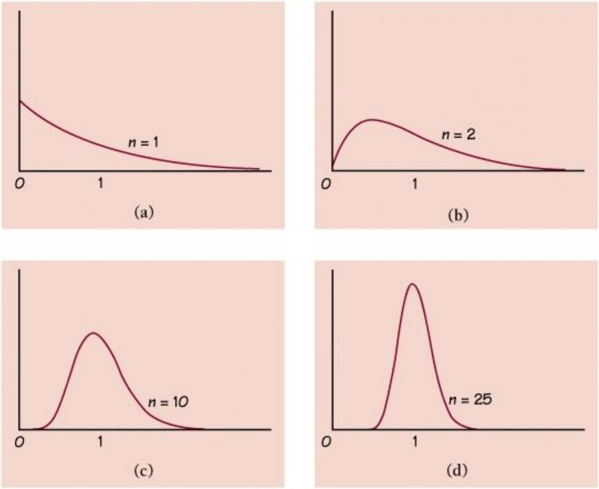
SAMLING DISTRIBUTION OF A SAMPLE MEAN

Draw an SRS of size n from a population that has the normal distribution with mean μ and standard deviation σ. Then the sample mean x-bar HAS THE normal distribution N(μ, σ/$\sqrt{n}$) with mean μ and standard deviation σ/($\sqrt{n}$).

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population whatsoever with mean μ and standard deviation σ.
When n is large, the sampling distribution of the sample mean x-bar is close to the normal distribution N(μ,σ/$\sqrt{n}$)with mean μ and standard deviation σ/($\sqrt{n}$).

The Central Limit Theorem allows us to use normal probability calculations to answer questions about sample means from many observations even when the population distribution is not normal



The central limit theorem in action: the distribution of sample mean x-bar from a strongly nonnormal population becomes more normal as the sample size increases. (a) the distribution of 1 observation; (b) the distribution of x-bar for 2 observations; (c) the distribution of x-bar for 10 observations; (d) the distribution of x-bar for 25 observations.

Last but not least.....The law of large numbers

Draw observations at random from any population with finite mean μ. As the number of observations drawn increases the mean x-bar of the observed values gets closer and closer to μ.